Revisión de Supuestos Estadísticos e Introducción a Análisis Espacial

IAE - 2025

Ricardo Pasquini

27 de agosto de 2025

Repaso

- Estadístico T es la base del Test de Hipótesis

$$t = \frac{\hat{\beta} - \beta_0}{SE(\hat{\beta})} = \frac{\hat{\beta} - \beta_0}{\sqrt{\hat{\sigma}^2 (X'X)^{-1}}}$$

  Donde:
  - $\hat{\beta}$ es el coeficiente estimado
  - β_0 es el valor bajo la hipótesis nula
  - $SE(\hat{\beta})$ es el error estándar del coeficiente
  - $\hat{\sigma}^2$ es la varianza estimada de los errores
- Este estadístico tiene una distribución T-student. Sin embargo, este resultado se construye sobre varios supuestos:
  - Los errores tienen una varianza constante
  - Los errores son independientes
  - Los errores tienen una distribución Normal

Heteroscedasticidad

- Varianza no constante en los errores del modelo
  En el modelo homoscedástico (ideal):

  $$Var(\epsilon_i) = \sigma^2 \text{ para todo } i$$

  En presencia de heteroscedasticidad:

  $$Var(\epsilon_i) = \sigma_i^2 \text{ varía con } i$$

- La dispersión de los residuos varía según los valores de X o Y
- Puede surgir por:
  - Datos agrupados con diferentes niveles de variabilidad
  - Relaciones no lineales subyacentes
  - Errores de medición que varían con la magnitud de las variables

Heteroscedasticidad - Visualización

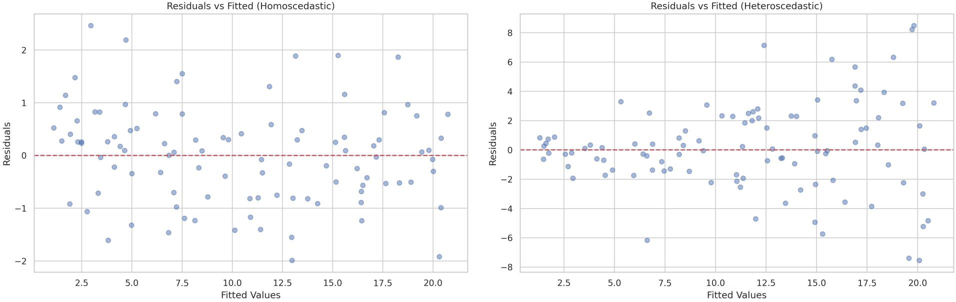

Figura: Residuos vs. Valores Ajustados: Homoscedástico (izq.) vs. Heteroscedástico (der.)

Consecuencias de la Heteroscedasticidad

- Estimadores OLS siguen siendo insesgados
- Errores estándar incorrectos
- Intuición del problema:
  - En OLS, asumimos que cada observación tiene la misma varianza (σ^2)
  - Cuando hay heteroscedasticidad, algunas observaciones son más ruidosas"que otras
  - La fórmula tradicional del error estándar:

$$SE(\hat{\beta}) = \sqrt{\hat{\sigma}^2(X'X)^{-1}}$$

    asume incorrectamente que todas las observaciones tienen el mismo peso
- En realidad, deberíamos dar:
  - Menor peso a observaciones con alta varianza
  - Mayor peso a observaciones con baja varianza

Soluciones para la Heteroscedasticidad

- Uso de errores estándar robustos (HC0 - White)
  - En Python (statsmodels):
    ```
    import statsmodels.formula.api as smf
    model = smf.ols('precio ~ superficie + habitaciones', data=df)
    results = model.fit(cov_type='HC0')
    ```
  - En R:
    ```
    coeftest(model, vcov = vcovHC(model, type = "HC0"))
    ```
- Otras correcciones de errores estándar robustos:
  - HC1: Corrección adicional para muestras pequeñas
  - HC2: Considera el leverage de las observaciones (el leverage es una medida de la importancia de la observación particular en el fit del modelo)
  - HC3: Más conservador, mejor para muestras muy pequeñas

Soluciones para la Heteroscedasticidad (cont.)

- Transformación de variables:
  - Transformación logarítmica cuando la varianza aumenta con el nivel
- Modelado explícito de la varianza:
  - Mínimos Cuadrados Ponderados (WLS)
  - Se asignan pesos inversamente proporcionales a la varianza estimada
  - Requiere conocer o poder modelar la estructura de la varianza

Normalidad de los Residuos

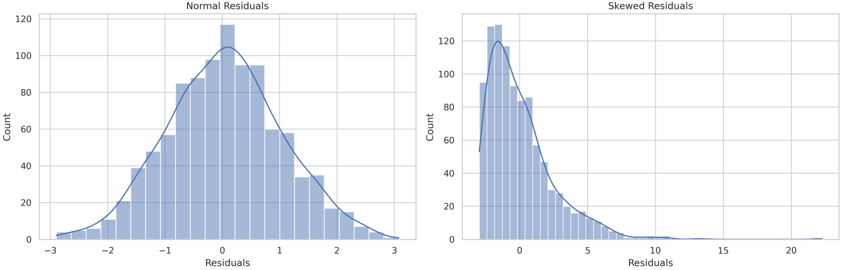

Figura: Distribución de Residuos: Normal vs. Sesgada

Normalidad de los Residuos - Q-Q Plot

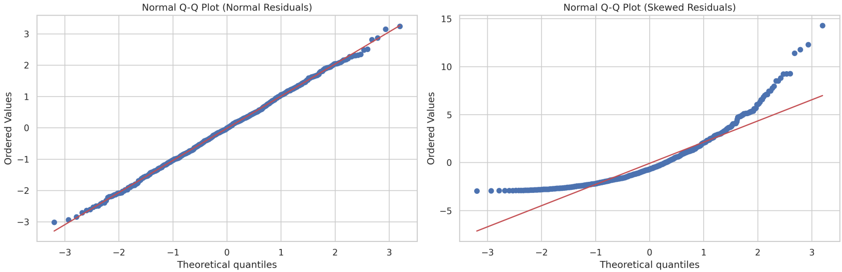

Figura: Q-Q Plots: Residuos Normales vs. Sesgados

Consecuencias e Importancia de la No-Normalidad

- Menos crucial cuando n es grande (Teorema asintótica)
- Las pruebas t y F son robustas con muestras grandes
- Mayor preocupación en muestras pequeñas
- Soluciones:
  - Transformaciones de la variable dependiente
    - Logarítmica
    - Raíz cuadrada
  - Identificar y tratar valores atípicos

Independencia de las Observaciones

- Tipos de Autocorrelación:
  - Temporal
    - Común en series de tiempo
    - Patrones sistemáticos en el tiempo
    - Observaciones cercanas en el tiempo están relacionadas
  - Espacial
    - Frecuente en estudios urbanos
    - "Todo está relacionado con todo, pero las cosas cercanas están más relacionadas entre sí"
    - Dependencia basada en la ubicación geográfica

Autocorrelación Temporal

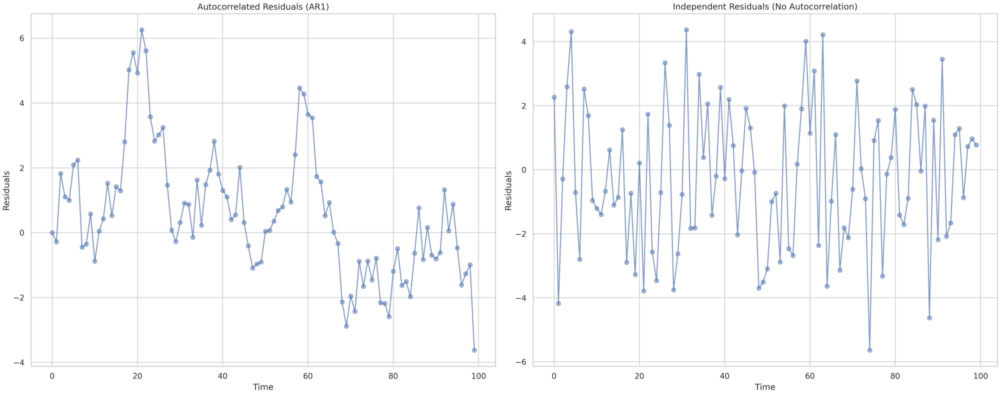

Figura: Comparación: Residuos con Autocorrelación Temporal (izq.) vs. Residuos Independientes (der.)

Autocorrelación Espacial

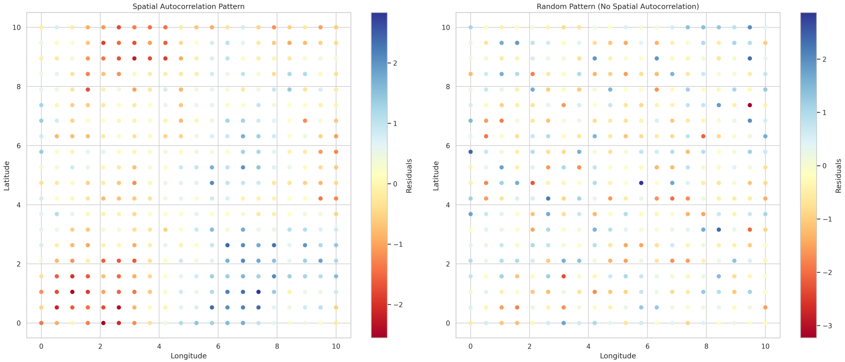

Figura: Residuos con Autocorrelación Espacial vs. Aleatorios

Análisis de Autocorrelación Espacial

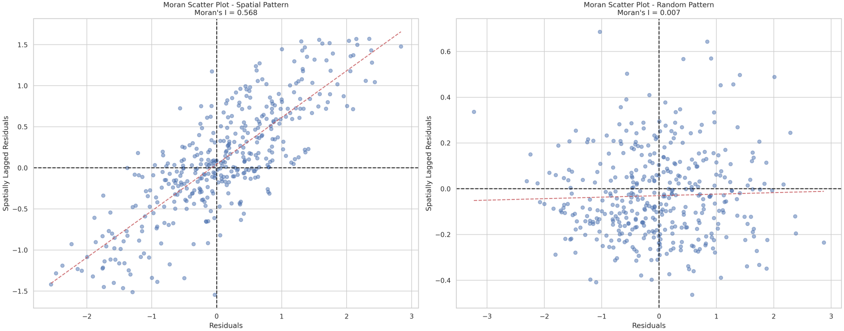

Figura: Diagramas de Moran: Patrón Espacial vs. Aleatorio

Consecuencias de la Autocorrelación

- Estimadores OLS ineficientes
- Errores estándar sesgados: cuando hay autocorrelación positiva, los errores estándar son menores que los reales (llevando a sobre-detectar significancia). Cuando hay autocorrelación negativa, los errores estándar son mayores que los reales (llevando a que no se detecte la significancia).
- Inferencia estadística no válida
- Predicciones subóptimas
- Soluciones:
  - Modelos con rezagos temporales (AR, MA, ARIMA)
  - Modelos espaciales
    - Rezago espacial (Spatial Lag)
    - Error espacial (Spatial Error)
    - Modelos mixtos espaciales
  - Corrección de errores estándar por clusters

Análisis Espacial

- ¿Qué hace que un análisis sea espacial?
  - Tanto la ubicación como los atributos importan
  - Los resultados cambian cuando cambia la disposición espacial
  - Contrasta con el análisis no espacial donde la ubicación es irrelevante
- Dimensiones de la econometría espacial:
  - Especificar la estructura de la dependencia espacial
  - Testear su presencia
  - Estimar modelos con efectos espaciales
  - Precaución: Los efectos espaciales podrían no cambiar las conclusiones

Desafíos Clave en el Análisis Espacial

- Falacia Ecológica
  - No se puede inferir comportamiento individual de datos agregados
  - Se debe mantener el nivel apropiado de interpretación
- Problema de la Unidad de Área Modificable (MAUP)
  - Los resultados dependen de la escala de agregación
  - Los resultados dependen de la zonificación/agrupación de unidades
  - Ejemplo: Distritos electorales y gerrymandering
- Problema de Cambio de Soporte
  - Mezcla de datos de diferentes escalas espaciales
  - Ejemplo: Datos puntuales vs. datos de área
  - Soluciones: agregar a una escala común, interpolar

Matriz de Pesos Espaciales (W)

- Propósito: Formaliza las relaciones espaciales entre observaciones
- Características:
  - Matriz N x N (N = número de observaciones)
- Mayormente dispersa
  (99 $\% cerostípicamente) Elementosdiagonales = 0(sinauto - vecinos)$
- Tipos Comunes:
  - Basados en Contigüidad
    - Rook (bordes comunes)
    - Queen (bordes + esquinas)
    - Bishop (solo esquinas)
  - Basados en Distancia
    - Bandas de distancia
    - K vecinos más cercanos
    - Funciones de distancia

Estadístico I de Moran

- Fórmula:

$$I = \frac{N}{W} \frac{\sum_i \sum_j z_i z_j \cdot w_{i,j}}{\sum_i z_i^2}$$

  donde:
  - $z_i = y_i - \bar{y}$ (desviación de la media)
  - N es el número de unidades espaciales
  - W es la suma de pesos
- Limitaciones:
  - Funciona como estadístico de mala especificación
  - Detecta correlación espacial, pero también:
    - Heteroscedasticidad
    - No linealidad

Introducción a la Regresión Espacial

- Desafío clave:
  - No se pueden modelar todas las interacciones con datos de corte transversal
  - N observaciones pero N^2 interacciones potenciales
  - Solución: Simplificar usando matriz de pesos espaciales

Variables Espacialmente Rezagadas

- Concepto básico: Promedio ponderado de observaciones vecinas
- Características:
  - Similar al rezago distribuido en series de tiempo
  - Suaviza la variabilidad (menor varianza)
  - Excluye auto-vecinos (diagonal de W es cero)

Modelo de Rezago Espacial

- Incluye la variable dependiente espacialmente rezagada (WY)
- Forma básica:

$$y = \rho W y + X\beta + \epsilon$$

  donde:
  - ρ es el coeficiente de autocorrelación espacial
  - Wy es el rezago espacial de la variable dependiente
  - $X\beta$ son las variables explicativas y sus coeficientes
- Características:
  - Modela interacción espacial directa
  - Tiene efectos multiplicadores (bucles de retroalimentación)
  - Su omisión lleva a sesgo e ineficiencia